QUAESTIONES GEOGRAPHICAE • 6 • 1980

A MONTE CARLO STUDY OF CORRELATION COEFFICIENT ESTIMATION WITH SPATIALLY AUTOCORRELATED OBSERVATIONS

ROGER BIVAND

Adam Mickiewicz University, Institute of Geography, Poznań, Poland

ABSTRACT. The problems involved in estimating the product moment correlation coefficient with spatially autocorrelated observations are set out, and a Monte Carlo experiment to explore risks made in inferring from such estimates is described. A range of different spatial lattices were used to investigate the errors introduced into the standard error estimator of Fisher's transformation of the correlation coefficient. The results of the experiment demonstrate that inference from correlation coefficient estimates in the presence of spatial autocorrelation is hazardous.

Introduction

The range of correlation coefficients available for the estimation of relationships between variables on the basis of sets of observations is wide. They share the property of varying between plus and minus unity, with zero representing the absence of a relationship between a given pair of variables. In the case of interval scaled observations, the Pearson product moment correlation coefficient is most frequently chosen; this does, however, require that the observations be relatively normal, since the coefficient is sensitive to major deviations from normality (Kowalski 1972, p. 11). After Siegel (1956, p. 18 - 34), one could reach for a non-parametric alternative, in the hope of getting an unbiased estimate of the coefficient. Unfortunately, one requirement shared by all correlation coefficients is that the observations from which they are calculated be independent.

In the case of a sample from a real or hypothetical population intended to estimate the unknown population value of a correlation coefficient, its null distribution may be altered (cf. Student 1914, Yule 1926). Unwin and Hepple conclude that "since the null distribution varies with each pair of series and their respective autocorrelations, ... the inferential analysis of correlation matrices, and associated multivariate statistics, such as factor analysis, becomes hazardous" (1974, p. 220). These considerations apply, of course, to series of observations in time or in space. Spatial autocorrelation is defined in the following way: "If the presence of some quality in a county of a country makes its presence in neighbouring counties morē or less likely, we say that the phenomenon exhibits spatial autocorrelation" (Cliff and Ord 1973, p. 1). Since individual observations convey information about their neighbours, the number of degrees of freedom of a given set of observations may be reduced. This phenomenon has been expressed by Cliff and Ord: "In general, when similar values of the variable tend to cluster in space, we say that the variable exhibits positive spatial autocorrelation... In the more usual situation of positive spatial autocorrelation, we face the problem that an observation carries less information than an independent observation, since it is partly predictable from neighbouring observations" (1975a, p. 725).

The problem considered below concerns the estimation of the product moment correlation coefficient from series of observations which are positively spatially autocorrelated. An attempt is made to estimate the values of the coefficient, and their differences from the values expected with strictly independent observations, thus yielding empirical estimates of the null distributions of the coefficients, for a number of spatial lattices, and a range of autocorrelated series.

Estimation

When the correlation coefficient describing the relationship of two series is unknown, it is possible to estimate it without the further use of statistical significance tests. When the observations form a "sample" from a hypothetical population, it has been argued that hypothesis testing may be heuristically acceptable (Cliff 1973, for an opposing view, cf. Henkel 1976). Unfortunately, statistics estimating parameters using series of observations are known to be biased by autocorrelation; for example, the variance of a spatial series with positive spatial autocorrelation will be biased downwards using classical estimators (Unwin and Hepple 1974, p. 220, Cliff and Ord 1975a, b, p. 300, Bivand forthcoming). A further requirement concerns the choice of spatial scale, since it is known that spatial series can be altered or combined to yield very different estimates of the correlation between two variables (Yule and Kendal 1966, pp. 320 - 323, Openshaw 1977).

The product moment correlation coefficient is estimated as the covariance of two series of observations, x and y, divided by the product of their standard deviations:

$$\hat{r}=\frac{\hat{\mathrm{Cov}}(x,y)}{\hat{\sigma}_{(x)}\hat{\sigma}_{(y)}}, \tag{1}$$

where $\hat{\mathrm{Cov}}(x,y)=$

$$=(n-1)^{-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}),$$

and $\hat{\sigma}_{(x)}=\sqrt{(n-1)^{-1}\sum_{i=1}^{n}(x_i-\bar{x})^2},$

$$\hat{\sigma}_{(y)}=\sqrt{(n-1)^{-1}\sum_{i=1}^{n}(y_i-\bar{y})^2}$$

$$\bar{x}=n^{-1}\sum_{i=1}^{n}x_i,$$

$$\bar{y}=n^{-1}\sum_{i=1}^{n}y_i.$$

Where the parameters of the process governing the generation of a series are known, means and variances may be obtained using estimators given by Cliff and Ord (1975a, p. 727). The notation adopted here reserves ρ for the process parameter, leaving r as the product moment correlation coefficient in the population, and $\hat{r}$ as its estimator.

Given an estimate of a population parameter, one is interested in its variance as a guide to inference about the value of the parameter. The construction of a confidence interval about the estimate allows us to test whether, for example, a parameter value of zero could fall within it. If the distribution of the estimator is normal, the demarcation of the confidence interval is made easier (Wonnacott and Wonnacott 1972, pp. 141 - 147). However, since the values of the product moment correlation coefficient are bounded, the distribution of its estimator is non-normal, but may be normalized using Fisher's transformation (Yule and Kendall 1966, p. 497):

$$z=\frac{1}{2}\ln\frac{1+r}{1-r},\quad r=\tanh(z),$$

and $\hat{z}=\frac{1}{2}\ln\frac{1+\hat{r}}{1-\hat{r}},\quad \hat{r}=\tanh(z).$

The estimator of the square root of the variance of $\hat{z}$, the standard error, may be expressed as (David 1938, Kendall and Stuart 1963, pp. 390 - 391):

$$\hat{\sigma}_{1\,(\hat{z})}=\sqrt{(n-3)^{-1}}\tag{3}$$

or

$$\hat{\sigma}_{2\,(\hat{z})}=\sqrt{(n-1)^{-1}+(4+r_0)/(2(n-1)^2)},\tag{4}$$

where r_0 is the hypothesized parameter value. Estimator (3) is recommended for $n>50$, and (4) for $n>25$, however, only (3) is used here. As may be seen, both depend on n, indicating that bias will be introduced by positive spatial autocorrelation, resulting in the underestimation of the standard error. This would lead to a narrower confidence interval being used for inference than is in fact justified by the information carried by the observations. In order to investigate the bias in the standard error estimator empirically, a Monte Carlo experiment was undertaken.

Monte Carlo Study

The purpose of the study was to estimate the values of $\hat{z}$ for a number of lattices, together with their biases and root mean square errors (Johnston 1972, p. 408). A simulation procedure was developed using samp es of n observations from a bivariate norma distribution. Spatial autocorrelation was introduced using a first order autoregressive model as has become accepted practice (Cliff and Ord 1973, p. 146 - 147, 1975a, p. 730, Cliff, Martin and Ord 1974, p. 285, Martin 1974, p. 189):

$$x_i=\rho_{(x)}\sum_{j=1}^{n} w_{ij}x_j+\varepsilon_i,\tag{5}$$
$$y_i=\rho_{(y)}\sum_{j=1}^{n} w_{ij}y_j+\xi_i,\quad i=1,2,\dots,n,$$

or in matrix form:

$$\begin{aligned}x&=\delta_{(x)}Wx+\varepsilon,\\ y&=\rho_{(y)}Wy+\xi,\end{aligned}\tag{6}$$

where $\varepsilon_i,\ \xi_i\sim N(0,1)$.

Parameters $\rho_{(x)}$ and $\rho_{(y)}$ specify the overall degree of spatial autocorrelation in the lattice; w_{ij} is an element of the matrix W of scaled weights, where:

$$w_{ij}=w^*_{ij}/\sum_{j=1}^{n} w^*_{ij},\tag{7}$$

$$w^*_{ij}=\begin{cases}1 \text{ if } i \text{ and } j \text{ are contiguous neighbours,}\\ 0 \text{ otherwise.}\end{cases}$$

It is clear that x and y may be generated by:

$$\begin{aligned}x&=(I-\rho_{(x)}W)^{-1}\varepsilon,\\ y&=(I-\rho_{(y)}W)^{-1}\xi.\end{aligned}\tag{8}$$

Where ρ is greater than zero, the transformed variable will contain n positively spatially autocorrelated observations. Since the calculation of the inverse matrix $(I-\rho W)^{-1}$, is inconvenient using standard procedures, a special inversion routine was written to invert the matrix using the power series of ρW, since $(I-A)^{-1}=\sum_{k=0}^{\infty}A^k=I+A+A^2+\dots$, where $1>a_{ij}\geqslant 0$ (Hadley 1961, p. 118 - 119). The series converges rapidly, especially with small ρ, reducing the time needed for inverting such large $n \times n$ and sparse matrices.

The experiment was carried out in the following way:

1° Choose a lattice from the following:

A – $n=25$, $\beta=2.00$, 5×5 regular lattice mapped onto a torus rook's case,

B – $n=29$, $\beta=2.21$, powiats of the former Poznań voivodeship,

C – $n=33$, $\beta=2.32$, taxonomic economic regions of Poland (Kukliński and Najgrakowski 1976, p. 55),

D – $n=49$, $\beta=2.00$, 7×7 regular lattice mapped onto a torus, rook's case,

E – $n=49$, $\beta=2.52$, Polish voivodeships from 1th June 1975;

where $\beta=e/n$, e=number of joins in the spatial system;

2° Select the values of ρ to be used, here 0.0, 0.5, 0.7, 0.9;

3° Calculate the necessary matrices $(I-\rho W)^{-1}$ using the power series procedure;

4° Generate 100 samples of $[\varepsilon\xi]$ using the same random number stream from *symnorm*, a multivariate random number generator (Krzyśko, Stolarski and Caliński 1973). The parameter values for the means, covariance matrix, and for z were $\mu=(0\,0)'$, $\Sigma=\begin{bmatrix}1.0 & 0.5\\ 0.5 & 1.0\end{bmatrix}$, and $z=0.5493$;

5° Choose a pair of values of ρ as $\rho_{(x)}$ and $\rho_{(y)}$;

6° Calculate **x** and **y** using (8);

7° Estimate the means and variances of the series, their covariance, correlation coefficient, and $\hat{z}$, repeating 6° and 7° 100 times;

8° Calculate the mean $\bar{\hat{z}}=\frac{1}{100}\sum_{p=1}^{100}\hat{z}_p$, the mean bias

$$BIAS=\frac{1}{100}\sum_{p=1}^{100}(\hat{z}_p-z),$$

and the root mean square error

$$RMSE=\frac{1}{100}\sum_{p=1}^{100}(\hat{z}_p-z)^2\,;$$

9° Return to 5° and choose a new combination of values for $\rho_{(x)}$ and $\rho_{(y)}$;

10° Return to 1° and choose a new lattice.

Results

The results of the Monte Carlo experiment are tabulated by lattice, clearly indicating the influence of spatial autocorrelation

TABLES 1 - 5. RESULTS OF MONTE CARLO EXPERIMENT

LATTICE A

$n=25$ 5×5 $\beta=2.00$
$z=.5493$
$\hat{\sigma}_{1(\hat{z})}=.2132$

$\rho_{[x)}$	$\rho_{(y)}$	$\bar{\hat{z}}$	BIAS	RMSE
.0	.0	.5680	.0187	.2042
.0	.5	.5470	−.0023	.1919
.0	.7	.5275	−.0218	.1890
.0	.9	.4966	−.0527	.1929
.5	.5	.5633	.0140	.2092
.5	.7	.5596	.0103	.2190
.5	.9	.5453	−.0040	.2282
.7	.7	.5631	.0138	.2362
.7	.9	.5588	.0095	.2527
.9	.9	.5651	.0158	.2831

LATTICE B

$n=29$ Poznań $\beta=2.21$
$z=.5493$
$\sigma_{1(z} =.1961$

$\rho_{1(\hat{x})}$	$\rho_{(y)}$	$\bar{\hat{z}}$	BIAS	RMSE
.0	.0	.5686	.0193	.1965
.0	.5	.5485	−.0008	.1944
.0	.7	.5152	−.0341	.1957
.0	.9	.4471	−.1022	.2170
.5	.5	.5817	.0323	.2344
.5	.7	.5780	.0287	.2549
.5	.9	.5402	−.0091	.2779
.7	.7	.5887	.0394	.2871
.7	.9	.5795	.0302	.3378
.9	.9	.6013	.0520	.4173

LATTICE C

$n=33$ Taxonomic regions $\beta=2.32$
$z=.5493$
$\hat{\sigma}_{1(\hat{z})}=.1826$

$\rho_{(x)}$	$\rho_{(y)}$	$\bar{\hat{z}}$	BIAS	RMSE
.0	.0	.5463	−.0030	.1906
.0	.5	.5261	−.0232	.1888
.0	.7	.4923	−.0570	.1891
.0	.9	.4223	−.1270	.2116
.5	.5	.5571	.0078	.2146
.5	.7	.5483	−.0010	.2274
.5	.9	.5032	−.0461	.2432
.7	.7	.5613	.0120	.2063
.7	.9	.5400	−.0093	.2920
.9	.9	.5703	.0210	.3805

LATTICE D

$n=49$ 7×7 $\beta=2.00$
$z=.5493$
$\hat{\sigma}_{1(\hat{z})}=.1473$

$\rho_{(x)}$	$\rho_{(y)}$	$\bar{\hat{z}}$	BIAS	RMSE
.0	.0	.5587	.0094	.1590
.0	.5	.5391	−.0102	.1517
.0	.7	.5133	−.0360	.1521
.0	.9	.4628	−.0865	.1677
.5	.5	.5688	.0195	.1712
.5	.7	.5659	.0166	.1809
.5	.9	.5409	−.0084	.1941
.7	.7	.5781	.0288	.1970
.7	.9	.5729	.0236	.2235
.9	.9	.5974	.0481	.2628

LATTICE E

$n=49$ Voivodeships $\beta=2.52$
$z=.5493$
$\hat{\sigma}_{1(\hat{z})}=.1473$

$\rho_{(x)}$	$\rho_{(y)}$	$\bar{\hat{x}}$	BIAS	RMSE
.0	.0	.5587	.0094	.15090
.0	.5	.5404	−.0089	.1526
.0	.7	.5073	−.0420	.1552
.0	.9	.4339	−.1154	.1880
.5	.5	.5680	.0187	.1760
.5	.7	.5638	.0145	.1907
.5	.9	.5212	−.0281	.2136
.7	.7	.5748	.0255	.2184
.7	.9	.5602	.0109	.2569
.9	.9	.5868	.0375	.3270

on the estimates of the standard error of $\hat{z}$, shown in the columns of root mean square error values. It is obvious that the rejection of a null hypothesis that the parameter $z=0.0$ may occur through the use of the biased estimator $\hat{\sigma}_{1(\hat{z})}$ where the autocorrelation of the spatial series is marked. Bias in the estimator $\hat{z}$ seems to stem rather from differences in the processes generating the series of observations.

A further point of interest in the bias of the standard error estimator is that it appears to vary from lattice to lattice, confirming the view expressed by Unwin and Hepple cited above. For irregular lattices with both series strongly autocorrelated, the width of the confidence interval given using the estimator $\hat{\sigma}^{1}_{(\hat{z})}$ would seem to be about half that necessary to test the null hypothesis with an approximately correct level of significance. For the Polish voivodeships lattice, $n=49$, $\beta=2.52$, $\hat{\sigma}_{1(\hat{z})}=0.1473$, $\text{RMSE}_{0.9},\ _{0.9}=0.3270$, and $\alpha=0.05$, the confidence interval would be $\hat{z}\pm1.96\times\times0.1473=\hat{z}\pm0.2887$, where the confidence interval corresponding to a 5% significance level using the estimated standard error is in fact $\hat{z}\pm1.96\times0.3270=\hat{z}\pm0.6509$. The confidence interval using $\hat{\sigma}_{1(\hat{z})}$ here thus represents a significance level not of 5% but of 37%. This seems to demonstrate conclusively the dangers of inferring from the product moment correlation coefficient when the observations may be spatially autocorrelated.

Dr. Roger Bivand, Institute of Geography. Adam Mickiewicz University, Fredry 10, 61-701 Poznań, Poland

References

Bivand R., forthcoming: Autokorelacja przestrzenna a metody analizy statystycznej w geografii. In: *Chojnicki Z.* (ed.) *Analiza regresji w geografii*, Warszawa.

Cliff A., 1973: A note on statistical hypothesis testing. *Area* 5, p. 240.

Cliff A., Martin R., Ord J. K., 1974: Evaluating the friction of the distance parameter in gravity models. *Regional Studies* 8, pp. 281 - 286.

Cliff A., Ord J. K., 1973: Spatial autocorrelation. London.

Cliff A., Ord J. K., 1975a. The comparison of means when samples consist of spatially autocorrelated observations. *Environment and Planning* A7, pp. 725 - 734.

Cliff A., Ord J. K., 1975b: Model building and the analysis of spatial pattern in human geography. *Journal of the Royal Statistical Society* B37, pp. 297 - 348.

David F. N., 1938: Tables of the ordinates and probability integral of the distribution of the correlation coefficient in small samples. London.

Hadley G., 1961: Linear algebra. Reading, Mass.

Henkel R. E., 1976: Tests of significance. *Sage University Paper series on quantitative applications in the social sciences* 4, Beverly Hills-London.

Johnston J., 1972: Econometric methods. New York.

Kendall M. G., Stuart A., 1963: The advanced theory of statistics, Vol. I. London.

Kowalski C., 1972: On the effects of non-normality on the distribution of the sample product moment correlation coefficient. *Applied Statistics* 21, pp. 1 - 12.

Krzyśko M., Stolarski P., Caliński T., 1973: Symulacja wielowymiarowego rozkładu normalnego. *Roczniki Akademii Rolniczej w Poznaniu* 64, pp.153 - 160, ABS-20.

Kukliński A., Najgrakowski M., 1976: Struktura procesów inwestycyjnych a rozwój regionalny. *Przegl. Geogr.* 48, pp. 51 - 60.

Martin R., 1974: On autocorrelation, bias and the use of first spatial differences in regression analysis. *Area* 6, pp. 185 - 194.

Openshaw S., 1977: A geographical solution to scale and aggregation problems in region--building, partitioning and spatial modelling. *Transactions*, IBG N.S.2, pp. 459 - 472.

Siegel S., 1956: Nonparametric statistics for the behavioral sciences. New York.

Student, 1914: The elimination of spurious correlation due to position in time or space. *Biometrika* 10, pp. 179 - 180.

Unwin D., Hepple L., 1974: The statistical analysis of spatial series. *The Statistician* 23, pp. 211-227.

Wonnacott T., Wonnacott R., 1972: Introductory statistics, New York.

Yule G. U., Kendal M. G., 1966: Wstęp do teorii statystyki. Warszawa.

Yule G. U., 1926: Why do we sometimes get nonsense-correlations between time-series? — a study in sampling and the nature of time series. *Journal of the Royal Statistical Society* 89, pp. 1 - 69.